DOCUMENT RESUME

ED 308 249 TM 013 669

AUTHOR van der Linden, Wim J.
TITLE Optimizing Incomplete Sample Designs for Item Response Model Parameters. Research Report 88-5.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Apr 88
NOTE 35p.
AVAILABLE FROM Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Identification; Computer Assisted Testing; Elementary Education; Elementary School Students; Equations (Mathematics); Foreign Countries; Grade 7; *Latent Trait Theory; Mathematical Models; Mathematics Achievement; *Maximum Likelihood Statistics; *Sampling; Statistical Analysis
IDENTIFIERS *Incomplete Sample Design; Linear Models; Netherlands; Optimization; *Parameter Identification

ABSTRACT

Several models for optimizing incomplete sample designs with respect to information on the item parameters are presented. The following cases are considered: (1) known ability parameters; (2) unknown ability parameters; (3) item sets with multiple ability scales; and (4) response models with multiple item parameters. The models are able to cope with hierarchical structures in the population of examinees as well as the domain of content, and allow for practical constraints with respect to such items as test content, curricular differences between groups, or time available for item administration. An example with test data from a national assessment study illustrates the use of the models. This methodology was applied to an imagined third study of the Dutch part of the Second Mathematics Study of the International Association for the Evaluation of Educational Achievement for the three subject areas of Geometry, Algebra, and Arithmetic for a sample of 400 seventh graders. The LANDO computer program was used to solve the models, illustrating their utility. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED308249

Optimizing Incomplete Sample Designs for Item Response Model Parameters

Research Report 88-5

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. NELISSEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Wim J. van der Linden

department of EDUCATION

University of Twente

Division of Educational Measurement and Data Analysis

TM013669

Colofon:
Typing: Mevr. L.A.M. Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

Optimizing Incomplete Sample Designs for Item Response Model Parameters

Wim J. van der Linden

Optimizing incomplete sample designs for item response model parameters / Wim J. van der Linden – Enschede : University of Twente, Department of Education, April, 1988. – 28 pages

Abstract

Several models for optimizing incomplete sample designs with respect to information on the item parameters are presented. The following cases are considered: (1) known ability parameters; (2) unknown ability parameters; (3) item sets with multiple ability scales; and (4) response models with multiple item parameters. The models are able to cope with hierarchical structures in the population of examinees as well as the domain of content, and allow for practical constraints with respect to, e.g., test content, curricular differences between groups, or time available for item administration. An example with test data from a national assessment study illustrates the use of the models.

Keywords: Item Response Theory; Maximum-likelihood Estimation; Incomplete Sample; Optimal Experiment; Linear Programming

Optimizing Incomplete Sample Designs for Estimating Item Response Model Parameters

It is well known in item response theory that the accuracy of the maximum-likelihood estimators of the item parameters depends on the abilities of the examinees in the sample. This fact can easily be demonstrated for the Rasch model. Suppose item $i = 1, \ldots, I$ can be characterized by a scalar parameter δ_i representing its difficulty, whereas the scalar parameter θ_j measures the ability of examinee $j = 1, \ldots, J$ to solve the items. For stochastic responses $A_{ij} \in \{0, 1\}$, the model gives the probability of event $\{A_{ij} = 1\}$ as

$$(1) \qquad P(A_{ij} = 1) = [1 + \exp(\delta_i - \theta_j)]^{-1},$$

with $-\infty < \delta_i, \theta_j < +\infty$ (Rasch, 1980). If $j = 1, \ldots, J$ and independent responses from all the examinees are used to estimate δ_i, then for $J \to \infty$ the asymptotic variance of the maximum-likelihood estimator is equal to the reciprocal of Fisher's information measure

$$(2) \qquad I(\delta_i; \theta_1, \ldots, \theta_J) = \sum_{j=1}^{J} p_i(\theta_j)[1 - p_i(\theta_j)]^{-1},$$

where $p_i(\theta_j) \equiv P(A_{ij} = 1)$.

Obviously,

$$(3) \qquad \max_{\theta_1,\ldots,\theta_J} I(\delta_i;\theta_1,\ldots,\theta_J) = I(\delta_i;\theta_1^*,\ldots,\theta_J^*)$$

with

$$\theta_j^* = \delta_i, \qquad j = 1, \ldots, J.$$

Thus, from the point of view of designing a sample for estimating a single item parameter, the case of all examinee parameters equal to the (unknown) item parameter would be ideal.

A slightly more complicated problem is the one of designing an optimal sample for the simultaneous estimation of I item parameters with possibly different values. Then for each item parameter each examinee has a different contribution to the information in the sample. A moment's reflection reveals that now the optimal distribution of examinees in the sample is a function of the desired distribution of information over the item parameters. Hence, this distribution has to be specified before the sample can be designed. The following model gives a solution to the problem. Let r_i be the relative amount of information about item parameter δ_i wanted from the sample. Let x_k be the unknown sample frequency of examinees with parameter value θ_k, $k = 1, \ldots, K$, where the grouping of examinees can be made arbitrarily fine. Finally, suppose the response variables A_{ij} of different examinees are independent. Then, because of the additivity of Fisher's measure, the optimal

distribution of examinees is the value of $(x_1, \ldots, x_k)$ solving the following integer programming problem:

$$(4) \qquad \text{maximize } \sum_{k=1}^{K} I(\delta_{i_o};\theta_k)x_k$$

subject to

$$(5) \qquad r_1^{-1} \sum_{k=1}^{K} I(\delta_1;\theta_k)x_k = \ldots = r_I^{-1} \sum_{k=1}^{K} I(\delta_I;\theta_k)x_k,$$

$$(6) \qquad \sum_{k=1}^{K} x_k = J,$$

$$(7) \qquad x_k = 0, 1, \ldots, J, \qquad k = 1, \ldots, K.$$

The idea underlying this formalization is to maximize the information for an arbitrary item i_o under the condition that the distribution of information over all item parameters must be proportional to $(r_1, \ldots, r_I)$. As a consequence, the distribution of the J examinees is chosen such that the relative distribution of information as specified by $(r_1, \ldots, r_I)$ is "blown up as far as possible". Algorithms for solving (4) to (7) can be found in the integer programming literature (e.g., Wagner, 1975, chap. 13) and are amply available in computer code.

The above problem was given only as a theoretical exercise; in practice, the item parameters are unknown, and for finite values of I an exact solution to the equality constraint in (5) is unlikely to exist. However, before

giving a practical solution based on the same idea, it is observed that the problem is complicated somewhat further if the assumption that each examinee gets the same set of items should be abandoned. Such incomplete samples occur in situations where the number of items is too large to administer all items to each examinee, for example, when large item banks or sets of parallel tests have to be constructed.

It is the purpose of the present paper to give some optimization models for incomplete sample designs. For these models to be practical, it is necessary that, unlike the above example, the item parameters need not be known beforehand. Several cases are considered. In the first case, the assumption of known item parameter values is dropped. The primary goal of presenting this case is to introduce the methodology and to show how it applies to the present problem. In the next case the assumption of known values for the ability parameters is dropped as well. The results are then generalized: first to the case of multiple ability scales and next to the case of models with multiple item parameters. An illustrative example concludes the paper. Following the practice of survey sampling (Cassel, Särndal, & Wretman, 1977), a distinction between stochastic and deterministic designs is made. All results are formulated as linear programming models. Linear programming was applied earlier in test theory to the problem of test construction: first by Theunissen (1985), and subsequently by Adema (1988), Boekkooi-Timminga (1987), Theunissen (1986), Theunissen and

Verstralen (1986), van der Linden (1987), van der Linden and Adema (1988), and van der Linden and Boekkooi-Timminga (1988, in press). A sequential solution to the problem of sample optimization not based on a linear programming approach is given in van der Linden and Eggen (1986).

General Shape of the Sample

Mostly, the design of the sample for calibrating a large item bank has to meet several conditions. The following types of conditions can be distinguished.

First, the domain of content for which the items are constructed may have a structure of problems nested in topics. The topics, in turn, may display a hierarchical structure of specific topics nested in broader ones. Also, different groups of topics usually require different abilities to solve the problems. Although the classification of items into topics is only a matter of judgment by experts, the ability structure enters the design as an empirical hypothesis to be tested on response data.

Second, the population of examinees usually also has a hierarchical structure, with examinees nested in classes, schools, districts, and so on. In addition, social or geographical factors may play a role. In the practice of item calibration, group-based designs are common. They reduce the logistic efforts involved in administering the items and leave units like school classes intact.

Third, various practical conditions may exist, for example, with respect to the composition of the tests, curricular differences between schools, or the time available for item administration. In order to avoid the problem of nonresponse, the sample should be designed to meet such conditions as well.

The general shape of the sample considered in the paper is given in Figure 1. The examinees in the operational

---

Insert Figure 1 about here

---

population are denoted by $j = 1, \ldots, J$. Each examinee is nested in one of the groups $g = 1, \ldots, G$. The ability scales are denoted by $s = 1, \ldots, S$, where scale s consists of topics $t = 1, \ldots, T_s$ and topic t contains item $i = 1, \ldots, I_t$. Additional levels of groups and topics can easily be introduced by further partitioning the examinees and the items, but are ommitted here for convenience. The design variables x_{jits} and x_{gits} represent the decision on examinee j and group g with regard to item (i,t,s). Two possible types of designs are considered: (1) Stochastic designs, where the variables are probabilities of assigning items to examinees (or groups). The problem is to optimize the probabilities; once they are known, auxiliary experiments can be used to decide on the actual administration of the items. Sampling with a stochastic design is common in survey research

(Cassel, Särndal, & Wretman, 1977), but has remained unnoticed in IRT; (2) Deterministic designs, where the variables take only the vales one or zero, indicating whether or not the item is administered to the examinee or the group.

The strategy in the following part of the paper is to start with the simple case of one scale and one topic and then to generalize to the more complete case in Figure 1. Although the models are formulated either for examinees or groups as sampling units, they can easily be transformed to the other level.

Unknown Item Difficulties and Known Abilities

Suppose we have at our disposal J examinees with known values for their ability parameters. Such cases are not frequently met in practice, but may arise in a two-stage testing strategy where a short pretest is used to optimally assign items to examinees in the second stage.

Since the values of the item parameters are unknown, a target for the information in the sample over the scale of possible values of the parameter has to be given. For a fixed parameterization, let $\delta_{i1}, \ldots, \delta_{id}, \ldots, \delta_{iD}$ be the possible values of parameter δ_i for which the target is specified (the number of points can be made arbitrarily large). Two types of models are presented each based on a different target: (1) A relative target, denoted by a vector of nonnegative numbers $\rho_{i1}, \ldots, \rho_{iD}$. These prespecified numbers indicate the information the sample should have on item parameter δ_i if it

would take one of the values of $\delta_{i1}, \ldots, \delta_{iD}$ relative to the others; (2) An absolute target $\alpha_{i1}, \ldots, \alpha_{iD}$ specifying on a fixed scale the actual amount of information the sample should have for each of the values $\delta_{i1}, \ldots, \delta_{iD}$. The former only specifies the shape of the distribution of information (e.g., uniform, skewed to the left, etc.); the latter also specifies its size. The relative target is used in the following model (cf. van der Linden, 1986).

Model of Maximal Information

It is assumed that the sample size is fixed and that the objective is to maximize the information in the sample such that the shape of the target distribution is realized maximizing its size. Formally, this can be achieved by replacing $(\rho_{i1}, \ldots, \rho_{iD})$ by the vector of products $(\rho_{i1}y, \ldots, \rho_{iD}y)$, where y is a scale factor to be maximized.

The model is:

(8) maximize y

subject to

$$(9) \quad \sum_{j=1}^{J} I(\delta_{id};\theta_j)x_{ji} - \rho_{id}y \geq 0, \qquad i = 1, \ldots, I, \; d = 1, \ldots, D;$$

$$(10) \quad \sum_{i=1}^{I} x_{ji} \leq n_j, \qquad j = 1, \ldots, J;$$

$$(11) \qquad 0 \leq x_{ji} \leq 1, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J;$$

$$(12) \qquad y \geq 0.$$

In (8), y is maximized so that the lower bounds to $\sum_{j=1}^{J} I(\delta_{id};\theta_j)x_{ji}$ in (9) are maximal while still obeying the ratio's specified by $(\rho_{i1}, \ldots, \rho_{iD})$. The length of the test for examinee j is restricted to n_j in (10), and the constraints in (11) specify a design with stochastic variables. The model is linear in the design variables and can be solved by the simplex algorithm (e.g., Wagner, 1975, chap. 4). Also, for any set of constants $\rho_{id} \geq 0$ and $n_j \geq 0$ a feasible solution exists. This property has been achieved by replacing the equality constraints in (5) by the inequalities in (9). Further, because the lower bounds in (9) are maximized, the solution hardly exceeds these bounds. Therefore, in practice, (9) works as an equality constraint. For the same reason, the expected test length will reach the upper bound in (10).

It should be noted that the constants ρ_{id} and n_j allow specification of the results at the level of individual items and examinees. This is the most general formulation of the model. However, in practice the target for the information in the sample will be the same for all items, whereas the required test length will be the same for groups of examinees. This reduces the number of coefficients to be specified in the model.

Unknown Item Difficulties and Unknown Abilities

Suppose now the values of the ability parameters are not known. In such cases, however, some kind of prior knowledge about the distribution of the abilities in groups of examinees may be available. For example, if a new item bank has to be calibrated, norms for the old bank may be known; or, perhaps, in a national assessment study of educational achievements results from a previous study could be used.

Such information can be exploited to optimize the sample, provided one of the old items is added to the sample to equate the scale for the new items to the old one. Although the following models are formulated for empirical priors at group level, the same type of models can be formulated for subjective priors at the level of individual examinees if one is ready to accept the assumption of exchangeability of examinees.

Model of Maximal Expected Information

Let $\Gamma_g(\theta)$ be the prior distribution for the population from which group g can be sampled with size N_g. The expected information on item parameter δ in group g is equal to

$$(13) \qquad E_g[I(\delta_{id};\theta)] = N_g \int I(\delta_{id};\theta)d\Gamma_g(\theta).$$

The model of maximal expected information follows on substituting (13) for $I(\delta_{id};\theta)$ in (8) through (12), replacing the design variables x_{ji} by x_{gi}.

Model of Minimal Sample Size

If the objective is not to maximize the expected information, but to minimize the sample size for an absolute target, additional decision variables are needed to include group g in the sample:

$$(14) \qquad z_g \equiv \begin{cases} 1 & \text{group g in the sample} \\ 0 & \text{otherwise} \end{cases} \qquad g = 1, \ldots, G.$$

The model is as follows:

$$(15) \qquad \text{minimize } \sum_{g=1}^{G} z_g$$

subject to

$$(16) \qquad \sum_{g=1}^{G} E_g[I(\delta_{id};\theta)]x_{gi} \geq \alpha_{id}. \qquad i = 1, \ldots, I, \quad d = 1, \ldots, D;$$

$$(17) \qquad n_g^{(1)} z_g - \sum_{i=1}^{I} x_{gi} \leq 0. \qquad g = 1, \ldots, G;$$

$$(18) \qquad \sum_{i=1}^{I} x_{gi} - n_g^{(2)} z_g \leq 0. \qquad g = 1, \ldots, G;$$

$$(19) \qquad z_g,\ x_{gi} \in \{0,\ 1\}, \qquad i = 1, \ldots, I, \quad g = 1, \ldots, G.$$

The objective function in (15) minimizes the number of groups, while (16) guarantees that the information in the sample is not below target. The latter also implies that each item is assigned to at least one group. The constraints in (17) and (18) serve a double goal: they set lower and upper bounds $n_g^{(1)}$ and $n_g^{(2)}$ to the length of the test for group g at the same time preventing the possibility of excluding g from the sample while assigning items to it. Because the objective is to minimize the sample size, for most groups only the upper bound in (18) is effective. In (19), deterministic design variables are specified. For this type of variable, a branch-and-bound algorithm can be used to solve the model (e.g., Wagner, 1975, chap. 13). Adema (1988) gives a modified branch-and-bound procedure that solves large-scale zero-one programming problems in realistic time. If the design variables are chosen to be stochastic, (17) and (18) set lower and upper bounds to the expected test length.

When specifying $n_g^{(1)}$, it should hold, of course, that $\sum_{g=1}^{G} n_g^{(1)} \geq I$. If this condition is satisfied, the model has a feasible solution if the target is set not too high relative to the maximum of $\sum_{g=1}^{G} z_g$.

Multiple Ability Scales

The above models are also appropriate for the case of multiple ability scales.

Suppose that S abilities are necessary to cover the complete set of items, and that each item requires exactly one of these abilities. Further, for each of the items in set $s = 1, \ldots, S$ the response model in (1) holds, but separate likelihoods are necessary to estimate the item parameters in different sets. Let I_s denote the set of indices of the items in s, while ability θ_s is needed to solve an item in s.

The only thing necessary to apply the above models to the case of multiple abilities is to take account of scale differences between the item sets. For the case of known abilities, the item parameter scales in the different sets of items are automatically fixed by their ability scales. If empirical priors are used, the scales of the item parameters have to be equated to the tests from which the priors were extracted. As already noted, this can be done by adding for each set an old item to the sample.

Suppose the interest is in the group-based model for maximal expected information, that is, in (8) through (12) with (9) replaced by (13). It should be noted that for this model, the effect of one numerical unit of the objective function depends on the item parameter scales chosen. This should be taken into account when specifying the targets for the separate item sets relative to each other. Let ρ_{id}, $d = 1, \ldots, D$, be the target for item $i \in I_s$, $s = 1, \ldots, S$,

and let $\Gamma_{gs}(\theta_s)$ be the prior for ability θ_s in group g. The model is now:

(20) maximize y

subject to

(21) $\sum_{g=1}^{G} (N_g \int I(\delta_{id};\theta_s)d\Gamma_{gs}(\theta_s)])x_{gi} - \rho_{id}y \geq 0,$
$i \in I_s,$
$s = 1, \ldots, S,$
$d = 1, \ldots, D;$

(22) $\sum_{i=1}^{I} x_{gi} \leq n_g,$ $g = 1, \ldots, G;$

(23) $0 \leq x_{gi} \leq 1,$ $i = 1, \ldots, I,$
$s = 1, \ldots, S,$
$g = 1, \ldots, G;$

(24) $y \geq 0.$

Multiple Item Parameters

So far, the Rasch model in (1) has been assumed to explain the probabilities of correct item responses. However, the only role it played was in the coefficients of constraints like (9), (13), (16), or (21); the structure of the optimization models was independent of the specific form of

the IRT model. In fact, only the number of item parameters counts, and the above models can be used for any one-parameter IRT model. The question arises how these models look like for IRT models with more than one item parameter.

For the sake of illustration, the two-parameter logistic model is adopted:

$$(25) \qquad p_i(\theta_j) = (1 + \exp[a_i(b_i - \theta_j)])^{-1},$$

with $-\infty < b_i, \theta_j < +\infty$ and $a_i > 0$, where a_i and b_i are the discriminating power and difficulty of item i, respectively (e.g., Hambleton & Swaminathan, 1985, sect. 3.3.2; Lord, 1980). For Fisher's information measure, it follows from (25) that

$$(26) \qquad I^*(\pi_i;\theta_1,\ldots,\theta_J) = \sum_{j=1}^{J} (\partial p_i(\theta_j)/\partial \pi_i)^2 (p_i(\theta_j)[1-p_i(\theta_j)])^{-1},$$

$$\pi_i = a_i, b_i.$$

The problem is to design the sample simultaneously with respect to information on parameters a_i and b_i. How this can be achieved by a simple reinterpretation of the above models is shown for the model in (15) through (19).

Let $a_{i1}, \ldots, a_{iv}, \ldots, a_{iV}$ and $b_{i1}, \ldots, b_{iw}, \ldots, b_{iW}$ be the possible values of parameters a_i and b_i for which absolute target values for the expected information are

specified. As in (26), let π_i run over (a_i, b_i); also let p run over (v, w). The model in (15) through (19) generalizes to the case of two item parameters, if (16) is replaced by

$$(27) \qquad \sum_{g=1}^{G} E_g[I^*(\pi_{ip};\theta)]x_{gi} \geq \alpha_{ip}, \qquad \begin{array}{l} i = 1, \ldots, I, \\ p = v, w, \\ v = 1, \ldots, V, \\ w = 1, \ldots, W. \end{array}$$

The only difference is the doubling of the number of constraints in (27) and a change of coefficients; the structure of the model remains the same.

Again, it should be noted that the targets for different parameters are set on different scales. If the difficulty scale is fixed by known abilities or the use of a known prior distribution, the scale of the discriminating power parameters is also fixed, but with reciprocal unit.

From the above, the generalization to more than two item parameters is obvious.

Practical Conditions on the Tests

As noted earlier, several practical conditions on the tests for the various groups in the sample design may be in force. Such conditions can be inserted in the optimization models given here without destroying their properties, provided they

can be formulated as linear (in)equalities. Some examples are given.

Test content. If for some groups of examinees the number of items on topic t are not allowed to be larger than n_{gt}, constraints like (10) should be replaced by

$$(28) \qquad \sum_{i \in I_t} x_{gi} \leq n_{gt}, \qquad \text{for some } g \text{ and } t.$$

Because in the models for maximal information (28) is likely to work as an equality constraint, the composition of the test is fixed if $\sum_{t=1}^{T} n_{gt} = n_g$.

Curricular fit. Suppose different groups have experienced different curricula with respect to the domain of content covered by the test items. If topic t has not been taught to group g, its items can be suppressed by setting $n_{gt} = 0$ in (28). In addition, it is possible to suppress individual items by

$$(29) \qquad x_{gi} = 0, \qquad \text{for some } i.$$

Test security. For reasons of test security it may be desirable to prevent some items from administering to more than v groups. This is achieved by the following linear constraint:

$$(30) \qquad \sum_{g=1}^{G} x_{gi} \leq v, \qquad \text{for some } i.$$

Simultaneous inclusion of groups. If groups with different priors belong to the same organization and a model of minimal sample size is chosen, it may be useful to make the same decision about inclusion in the sample for these groups:

$$z_g = z_{g'}, \qquad \text{for some } (g, g'). \tag{31}$$

However, the same effect can be obtained by replacing the priors for such groups by mixtures.

The above constraints are only a small selection from the possibilities. In principle, the same constraints may be needed as when designing customized tests from a calibrated item bank. A more complete treatment of the latter is given in van der Linden and Boekkooi-Timminga (in press).

Example

At the end of the seventies, the Dutch part of the Second Mathematics Study of the International Association for the Evaluation of Educational Achievement (IEA) was conducted. For several subjects the mathematics achievements of grade-seven pupils were assessed and compared with results in other countries. Suppose a third study has been planned and the methodology in this paper has to be used to optimize the sampling design for calibrating the test items. As model for the item responses, the Rasch model in (1) is chosen. Three different subject areas are distinguished (Geometry, Algebra,

and Arithmetic), and achievements in three different types of secondary education are to be reported (LTE: Lower Technical Education; DSE: Domestic Science Education; MGE: Middle General Education).

For a sample of 400 examinees from the Second Mathematics Study the Rasch model with a normal distribution over the ability parameters was fitted to the data. The results are given in Table 1. The numbers of items were equal

---

Insert Table 1 about here

---

to 10 (Geometry), 9 (Algebra), and 11 (Arithmetic). All estimates of μ and σ were transformed to the same scale, which was obtained by fitting the model to the total set of items for the LTO examinees ($p = .428$).

The normal distributions in Table 1 were used as empirical priors in the model of maximal expected information in (8) through (13) with (9) replaced by (13). The sampling plan used in this example is given in Table 2. Since no items

---

Insert Table 2 about here

---

were assumed to deserve a special treatment, the same uniform target was set for all items (i.e., $\rho_{id} = 1$, $i = 1, \ldots, I$,

d = 1, ..., D). The estimated difficulties of the previous items varied between −.852 and 2.010; therefore, the target was set at the points $\delta_1 = -.50$, $\delta_2 = +.50$, and $\delta_3 = +1.50$. The design variables in the model were allowed to take values between zero and one.

The model was solved using a standard computer program for linear programming problems on a DEC2060 system (LANDO; Anthonisse, 1984). The result, which is just an array of numerical values for the 900 design variables, is not shown here. For future reference, it is only noted that 831 variables were equal to zero or one whereas 69 variables took values between zero and one. The CPU time needed was 28 mins. and 27 secs. In addition, the value of the objective function for the solution was obtained. It was equal to 71.228. Since this is a lower bound to Fisher's information, it follows that, uniformly in δ, an asymptotic standard error of estimate not larger than $(71.228)^{1/2}=.118$ can be expected.

Final Remarks

For the sample design in the above problem, the actual assignment of items to examinees could take place using a random experiment. However, other options are available.

First, the design variables in the model could be replaced by zero–one variables. Since the number of variables in the example generally is too large for a branch–and–bound algorithm, a heuristic as in Adema (1988) could be used. The heuristic is based on the principle that, for a maximization

problem, the value of the objective function for a solution with relaxed variables is an upper bound to the one for the zero–one problem. Hence, as the final solution it accepts the first feasible solution with an objective function value differing no more than a small percentage , 0.5%, say, from the upper bound. The heuristic has proven to solve large–scale problems in realistic time (Adema,1988).

Second, for larger numbers of decision variables than in the example, splitting a zero–one problem into smaller problems for which a branch–and–bound algorithm could be used is an attractive option. In the example, the design matrix could be partitioned along its rows, computing an optimal solution for subsets of groups. In this option, the model has not to be adapted but for the range of the group index. If an absolute target were to be used, targets of subproblems should allow for possible differences in numbers of subjects.

The final option for a model as in the example is rounding the solution of the relaxed problem. It is known that the number of decision variables in the solution with noninteger values is not larger than the number of constraints (Dantzig, 1957); in practice, it is considerably smaller. The number of noninteger values in the example was equal to 69. These values can be rounded optimally by a zero–one programming problem consisting of the previous problem with the other variables fixed at their values in the previous solution. The objective now is to approach the information on the item parameters in the this solution as close as possible.

References

Adema, J.J. (1988). A note on solving large-scale zero-one programming problems. Operations Research Letters, submitted for publication.

Anthonisse, J.M. (1984). Lando. Amsterdam: Centre for Mathematics and Computer Science CWI.

Cassel, C.M., Särndal, C.E., & Wretman, J.H. (1977). Foundations of inference in survey sampling. New York: Wiley.

Dantzig, G. (1957). Discrete-variable extremum problems. Operations research, 5, 266-277.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago, Ill.: The University of Chicago Press (2nd. edition).

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-390.

Theunissen, T.J.J.M., & Verstralen, H.H.F.M. (1986). Algoritmen voor het samenstellen van toetsen [Algorithms for constructing tests]. In W.J. van der Linden (Ed.), Moderne methoden voor toetsconstructie en -gebruik. Lisse: Swets & Zeitlinger.

van der Linden, W.J., & Adema, J.J. (1988). Algorithmic test design using classical item parameters. Applied Psychological Measurement, submitted for publication.

van der Linden, W.J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subtests method. Applied Psychological Measurement, 12, to appear.

van der Linden, W.J., & Boekkooi-Timminga, E. (in press). A maximin model for test design with practical constaints. Psychometrika.

van der Linden, W.J., & Eggen, T.J.H.M. (1986). An empirical Bayes approach to item banking. Applied Psychological Measurement, 10, 345-354.

Wagner, H.M. (1975). Principles of operations research. London: Prentice-Hall.

Table 1

Parameter estimates of the empirical priors for different types of schools and subject areas.

Subject Area	Parameter	Type of School .LTE	.DSE	.LGE
Geometry	μ	−.446	−1.263	.053
	σ	.809	.665	.920
	$p^{1)}$	.529	.577	.749
	$p^{*2)}$	.658	.237	.499
Algebra	μ	.446	−.440	1.657
	σ	.700	.704	.950
	p	.423	.670	.145
	p^{*}	.056	.474	.871
Arithmetic	μ	−.147	−.690	.649
	σ	.950	.512	.888
	p	.344	.223	.256
	p^{*}	.962	.666	.792

Notes: 1) probability of exceeding the result for the likelihood-ratio test on the Rasch model; 2) idem for assumption of normality.

Table 2

Numbers of groups and items and group sizes in the sample.

Types of School	No. of Groups	Group Size	Subject Area	No. of Items
LTE	4	32	Geometry	25
DSE	6	35	Algebra	15
MGE	5	28	Arithmetic	20

Scale	1	...	s	...	S
Topic		1 ...	t ...	T_s	
Item		1 ...	i ...	I_t	

Group	Examinee	
	1	
1		
	⋮	
⋮		
g	j	x_{jits}
		x_{gits}
⋮	⋮	
G		
	J	

Figure 1. General shape of a sample for estimating IRT parameters

Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

RR-87-1 R. Engelen, Semiparametric estimation in the Rasch model

RR-87-2 W.J. van der Linden (Ed.), IRT-based test construction

RR-87-3 R. Engelen, P. Thommassen, & W. Vervaat, Ignatov's theorem: A new and short proof

RR-87-4 E. van der Burg, & J. de Leeuw, Use of the multinomial jackknife and bootstrap in generalized nonlinear canonical correlation analysis

RR-87-5 H. Kelderman, Estimating a quasi-loglinear models for the Rasch table if the number of items is large

RR-87-6 R. Engelen, A review of different estimation procedures in the Rasch model

RR-87-7 D.L. Knol & J.M.F. ten Berge, Least-squares approximation of an improper by a proper correlation matrix using a semi-infinite convex program

RR-87-8 E. van der Burg & J. de Leeuw, Nonlinear canonical correlation analysis with k sets of variables

RR-87-9 W.J. van der Linden, Applications of decision theory to test-based decision making

RR-87-10 W.J. van der Linden & E. Boekkooi-Timminga, A maximin model for test design with practical constraints

RR–88–1 E. van der Burg & J. de Leeuw, Nonlinear redundancy analysis

RR–88–2 W.J. van der Linden & J.J. Adema, Algorithmic test design using classical item parameters

RR–88–3 E. Boekkooi–Timminga, A cluster–based method for test construction

RR–88–4 J.J. Adema, A note on solving large–scale zero–one programming problems

RR–88–5 W.J. van der Linden, Optimizing incomplete sample designs for item response model parameters

Research Reports can be obtained at costs from Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

department of
EDUCATION

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands